SYNTACTIC AND SEMANTIC PROBLEMS IN AUTOMATIC SENTENCE GENERATION

Kenneth E. Harper

A program for automatic sentence generation (ASG) has been developed in the linguistic group at The RAND Corporation, as a device for studying selected problems in syntax and distributional semantics. The general purpose of the study is to test certain linguistic assumptions by experimentation: what is the result when a computer program "composes" sentences on the basis of these assumptions? The assumptions, and their implementation in the computer program, are subject to modification, as deficiencies are observed in the output sentences. This trial-and-error procedure is presently at an early stage of development; the present paper is a brief, non-technical description of the procedure, and a report on some of the initial problems encountered, together with tentative solutions. The ultimate goal of the research is the generation of "paragraphs" (meaningful strings of sentences) as a contribution towards automatic abstracting.

1. The Sentence Generation Routine

The ASG program operates with Russian language materials, for the simple reason that the kind of data on which the linguistic assumptions are based is available only for Russian. The data is derived from the corpus of Russian physics text processed at The RAND Corporation (References 1, 2). In effect, the program described here deals with a sub-set of the Russian language found in these physics texts. The program may be said to consist of three parts: a glossary of words, a grammar, and a program for synthesizing sentences.

(i) The glossary is merely a list of 550 Russian words on magnetic tape; each entry consists of a "word number" (the word-identification number in the RAND physics glossary), (Reference 3) and a representative Russian form (not necessarily the canonical form). The glossary comes into operation only at the end of the sentence generation routine. Generally, the words in the glossary are used in a single sense in physics texts.

(ii) The grammar employed is a simplified dependency grammar for Russian. Its essential feature is that it is word-specific: each word in the glossary is provided a list of words with which it may combine in a dependency or governor relationship. The basic principle is that syntactic cooccurrence (e.g., the pairing of a given noun with a given adjective) is allowed only if this pairing is attested in the previously processed physics text. (An exception is made in the pairing of members of Semantic Classes, as explained below.)

The following information, available for each word in the glossary, may be said to constitute the grammar:

1) Word number
2) Part-of-speech code. Six such codes are used: VT (transitive verb), VI (intransitive verb), N (noun), A (adjective), DV (adverb), and DS (sentence adverb). A word may bear only one such code.
3) Semantic Class code (SC). Twenty such classes are recognized; all are distributionally formed. (References 4, 5)
4) Set of governing probabilities
5) Set of dependent probabilities
6) Coordinate probability

7) List of word numbers of governors
8) List of word numbers of dependents
9) List of word numbers of coordinates

Table 1 illustrates the meaning of governing probabilities. Reading across the top line in the table, we see that each VT has a probability of P_1 of governing some noun as subject and a probability of 1 of governing some noun as object. The VT has probabilities of P_2 and P_3 of governing an adverb and a sentence adverb respectively, and a probability of 0 of governing anything else. The table shows also that each VI must govern a noun as subject, and governs a DV and DS with probabilities P_4 and P_5 respectively; each noun governs another noun or an adjective with probabilities P_6 and P_7; a DS has a probability of 1 of governing a noun; adjectives and adverbs never govern.

Table 2 shows the meaning of dependent probabilities; here, reading across, the various probabilities of being governed by other parts-of-speech are shown for nouns, adjectives, adverbs, and sentence adverbs; verbs do not have governors.

The sets of governing probabilities for a word are associated with independent situations. For example, a transitive verb governs a noun as subject, a noun as object, an adverb and a sentence adverb independently. Therefore, in sentence generation, the decision to select a dependent type will be made without regard to any dependent types already selected for that governor. This is not the case, however, when selecting a governor for a particular word. The possibilities here are dependent events. An adverb, for example, must be governed by either a transitive or intransitive verb. Thus, the set of dependent probabilities for each word will add to one.

Probabilities for coordination are assigned for relatively few words. Since coordinate conjunctions are not represented as a word class, the assignment of a Russian form for "and" and "or" will be generated by the program (rather than by the glossary) in the sentence output stage.

As previously mentioned, each word is accompanied by a list of the word numbers that may function as its governors and dependents. Generally, these dependency pairs have been attested in physics text. However, a Semantic Class may also function as a word's governor or dependent; when this happens, the program is free to choose randomly any member of the SC in building the dependency pair. Many such pairs will not have been attested in text; the purpose here is to test the adequacy of the Semantic Classes in word combination. A complete list of the SC's cannot be given here; examples are (a) nouns that name physical properties ("height," "weight"), (b) verbs and nouns referring to a quantitative change ("to increase," "change"), and (c) the names of physical particles ("atom," "proton"). The classes, and their members, are purely experimental and are subject to modification.

(iii) The program for ASG (written in MAP for the IBM 7044) is intricately bound up with the grammar, but for purposes of discussion we may consider it separately. Essentially, the program has three functions. First, it restructures the grammar, in order to access it with minimal search time. (The grammar and the program are maintained in core storage.) Secondly, the program generates sentences; beginning at some arbitrary point, it proceeds, pair-wise, up and down the tree

Possible Dependents

Governor

	VT	VI	N		A	DV	DS
VT	0	0	S p_1	θ 1	0	p_2	p_3
VI	0	0	S 1	θ 0	0	p_4	p_5
N	0	0	p_6		p_7	0	0
A	0	0	0		0	0	0
DV	0	0	0		0	0	0
DS	0	0	1		0	0	0

Table 1

Governing Probabilities

Possible Governors

Dependent

	VT	VI	N	A	DV	DS
VT	0	0	0	0	0	0
VI	0	0	0	0	0	0
N	S p_1 / θ p_2	S p_5 / 0 0	p_8	0	0	p_9
A	0	0	1	0	0	0
DV	p_3	p_6	0	0	0	0
DS	p_4	p_7	0	0	0	0

Table 2

Dependent Probabilities

structure until a terminal point is reached. In the process, the algorithm provides that decisions of two kinds will be made at each node: (i) shall a dependent (or governor) be chosen for the word at this node, and (ii) if so, which word shall be chosen to complete the dependency pair? The grammar supplies the basic information necessary for making decisions; a pseudo-random number generator with uniform distribution is used in conjunction with the grammar when a choice exists. Finally, the program prints out the generated sentences in a fixed format. In their preliminary form, the sentences are merely strings of word numbers, together with associated data; glossary lookup then provides the transliterated Russian form associated with each word number. At present, the problem of morphology is bypassed: no attempt is made to supply the correct inflection of words in the sentence. The information necessary to carry out an inflection routine is available for nouns and adjectives (i.e., inflection for case and number); for verbs, person and number are specified, but tense is not. In the sentences discussed below, correct forms for nouns, verbs and adjectives have been supplied for reasons of clarity.

An example will perhaps serve to clarify the operation of the ASG program. In a greatly simplified way, the algorithm proceeds as follows.

(1) A starting point is chosen for the sentence. Possible starting points are transitive verb (VT), intransitive verb (VI), and noun (N). By random selection, VI is chosen for this sentence.

(2) Randomly, a particular VI is chosen from the list of VI's in the glossary. Here, we assume that word number 56410 is chosen.

(3) The possible dependent types of a VI are considered. (According to the grammar, verbs do not have governors, so that only dependents need be considered.)

(3.1) The grammar specifies that a VI must have a subject (the probability P is one for this pairing). The list of dependents serving as subject for word number 56410 is consulted, and an individual word is randomly chosen: word number 34550.

(3.2) A VI may have an adverb (DV) as dependent. For the verb in our sentence, the probability for this event is found to be .5. We assume that in this sentence the decision is made that a DV will be selected. The list of dependent DV's for word number 56410 is consulted, and an individual adverb is randomly selected: word number 14090.

(3.3) A VI may have a sentence adverb (DS) as a dependent. For our particular verb, the probability for this event is found to be .1. We assume that a decision in made not to select a DS dependent.

(4) Next, the dependents of the dependents of the verb are considered, but only with respect to their possible dependents (i.e., working down the dependency tree structure).

(4.1) The noun chosen as subject (word 34550, from 3.1 above) may have various dependents, with varying probabilities. Each of these is considered in turn. For sake of brevity, we assume that only one

is chosen in our sample sentence: an adjective. A particular adjective is randomly selected from the list of dependents for the noun: word number 63610.

(5) The adverb selected in 3.2 above has no dependents, according to the grammar, nor does the adjective selected in 4.1. At this point, the downward search for dependents is terminated, and the sentence is considered complete.

In form, the sentence is at this point a string of word numbers on a tape: 56410, 34550, 14090, 63610. Attached to each word number is data about its function in the sentence, its governor, the order in which it was selected, and, for nouns, an indication of grammatical number. Glossary lookup is then performed on this tape, and the transliterated Russian forms are printed out: uveličennoe otnošenie zametna ėta.

If fixed rules relating to word order and morphology are applied to to this string of forms, the sentence emerges as:

Ėto otnosenie zametno uveličivaetsja.

This ratio increases noticeably.

The foregoing is a drastically abbreviated description of the main steps in the ASG program. Sentences are presently generated at the rate of three per second; the addition of programming rules to account for morphology and word order would increase this time by an estimated ten percent. The program may, then, be considered as a practical, operational tool for research.

2. Discussion of Generated Sentences

At the present stage of development, the ASG program produces isolated sentences of varying degrees of complexity and "correctness." Since words in the glossary are limited in usage to one sense, and since semantic controls are guaranteed at least over pairs of words, a large number of sentences are quite acceptable. (The development of a context into which such sentences can be placed is, naturally, a far more difficult programming task.) The following are examples of "reasonable" sentences.

(1) Fejnman vyčislil integraly, s cel'ju opredelenija massy.
Feynman calculated the integrals in order to determine the mass.

(2) Rešenie zadač predlagaetsja v nastojaščej stat'e.
A solution of the problems is proposed in the present article.

(3) Ob"em kristallov izbytočnogo serebra bystro umen'šaetsja.
The volume of the crystals of excess silver rapidly decreases.

(4) Vozmožmost' sil'nogo vzaimodejstvija tela vnešnego istočnika privlekaet interes.
The possibility of the strong interaction of the body of the internal source is interesting.

Sentence (4) illustrates the approximate limit in number of levels

(six) for "reasonable" sentences in the present system. Additional levels can easily be generated, but only through the process of annexing genitive noun modifiers (i.e., English "of" phrases). Very few sentences have been generated with more than six levels, principally because of the drastically reduced probabilities of noun complementation at lower levels in the tree. Thus, most sentences are short. The chief obstacle to increasing sentence length (and complexity) is, however, the absence in the grammar of provisions for subordinate and coordinate clauses. Also inhibiting, from this point of view, is the absence of participles, prepositional phrases (beyond those used as adverbs), and pronouns. The price for adding any of these grammatical categories is increased complexity in the program. In an experimental situation, brevity and stylistic monotony can be tolerated.

Deficiencies in the generated sentences are of two main types: syntactic and semantic. Problems in both areas will be illustrated, although the line of demarcation is sometimes difficult to draw.

Syntactic problems are chiefly the result of inadequate complementation of nouns by adjectives or other nouns. It will be recalled that the grammar specifies for each noun the probability of its modification by an adjective or a genitive noun. These probabilities are assigned on the basis of the noun's behavior in text (Refernce 6). Since P is normally less than one for both kinds of modification, a given noun may frequently appear in a generated sentence without modification:

(5) Stepanov ustanovil teorii zadač.
Stepanov established theories of the problems.

(6) Formuly tipa ispol'zujutsja.
Formulae of the type are used.

(7) Razrjad issledovan pri rasčete.
The charge was studied in calculating.

(8) Proverka daet metod polučenija atoma.
Verification gives a method for obtaining the atom.

In (5), the noun phrase, "theories of the problems," appears to be ill-formed; the difficulty may not really be syntactic, since in an appropriate context the phrase may be nothing more than an instance of ellipsis. Nonetheless, specificity as to "what kind of problems" should be provided in the given sentence, or in preceding sentences. The unmodified use of "type" in sentence (6) is more difficult to justify on the basis of ellipsis; our tentative solution is to require either an adjective or noun modifier for words like "type," "kind," etc. (A modification in the program is necessitated here, since at present the selection of adjective and noun dependents is made independently.)

The nouns in (7) and (8) are strongly verbal, and appear to be deficient in complementation. This is particularly true of "rasčete" in (7), translated with the "-ing" form because of its verbal usage with the sentence adverb, "pri." It should be said that all sentence adverbs in our grammar have the property of conferring a verbal function upon noun dependents. (This is, of course, a very limited use

of sentence modifiers.) In order to maintain consistency in the grammar, it is clear that the program should be modified: the probability that a given noun will govern a genitive noun should equal one when the noun is itself governed by a sentence adverb.

In general, it is clear that a word's combining potential (i.e., its combining probabilities) may be affected by the syntactic environment in which it is placed. A generation program that does not take into account this possibility will be woefully inadequate. The problem is: which syntactic environments affect which words, and under which conditions? One use of the ASG program is to generate problems of this kind, and to test provisional solutions.

A second kind of syntactic problem arises when the grammatical number of nouns is inappropriate. (At present, the selection of number is made by the random number generator, operating on data in the grammar about the relative frequency of singular and plural in physics text.) Thus, in sentence (7) above, the combination "pri rasčetax" ("in calculations") could have easily been generated, since the noun, "rasčet," is frequently used in the plural. The strongly verbal nature of the noun in this environment, as noted above, makes the use of the singular noun almost imperative. (Deverbative nouns are almost never used in the plural when indicating a process.) The program should therefore be modified to require that a noun dependent of a sentence adverb be singular.

Two other instances of incorrect number in nouns may be mentioned. Nouns of the general classification, "abstract collective," require that genitive noun dependents be plural (unless the latter noun is rarely, or never, plural). Since the grammar contains no such specification, the following ill-formed sentence was generated:

(9) Čislo ètogo sloja otsutstvuet.
A number of this layer is absent.

Other nouns (deverbatives), and their corresponding transitive verbs, appear to require that the noun complement be plural.

(10) Izučenie stolknovenij atoma opublikovano v predyduščej rabote.
A study of the collisions of an atom was published in a preceding paper.

The solution to the problem is to modify the grammar so that "collison," will require the (subjective) genitive noun dependent in the plural. (A variation of this principle is "multiple complementation": "the collision of an atom with (and) another atom." Grammatical rules to account for this phenomenon are beyond the scope of the present grammar.)

We conclude that there are no major syntactic problems in the sentences so far generated, chiefly because the grammar is relatively primitive.

Semantic problems are more difficult to isolate. Again, it is sometimes possible that the absence of appropriate context is the chief cause of odd-sounding sentences.

A trivial kind of error is caused by the inappropriate repetition

of a word:

(11) Analogičnaja formula i formula zapisalis'.
An analogous formula and a formula were written.

It is easy to forbid the repetition of a word in a sentence; such a restriction, however, would be too severe (cf., for example, "We used formula 1 in deriving formula 2."). In general, it appears that recurring words tend to be used in different clauses of the sentence. Until the program is capable of producing sentences of more than one clause, the best strategy is probably to forbid word repetition.

In some sentences, the choice of adjective modifier appears logically inconsistent, or incompatible.

(12) Molekuly detal'no izučeny, s cel'ju raznogo izmerenija.
The molecules were studied in detail, for the purpose of a different measuring.

The issue here is not the lack of a complement for "measuring," but the use of "different" with the strongly verbal noun governor. In one sense, the difficulty may be syntactic: "different" would in Russian normally be used with a plural noun, whereas a singular noun is strongly indicated in the present context. The oddity of the construction is, thus, partly the result of conflicting "forces" in the adjective and the noun.

Some sentences show the loss of meaning that may easily result from the expansion of noun phrases by the addition of genitive noun modifiers.

(13) Uravnenie zakonov raspada èffekta približenij polučen.
The equation of the laws of the decay of the effect of approximation is obtained.

The problem here is apparently the use of "effect" in two different phrases: "the decay of the (e.g., photoelastic) effect" is the kind of phrase found in physics texts, as is "the effect of approximating (e.g., the energy)." The confusion in (13) may be explained by the fact that the former expression refers to a physical phenomenon, whereas the latter refers to an exercise of the mind; the two incompatible phrases are forced together by the fact that "effect" is common to both. The anomaly may also be explained, more simply, by the different lexical properties of "effect."

A similar problem may also arise in the use of two adjective modifiers in a noun phrase. Thus, in the construction, "adjective/noun of adjective/noun," varying degrees of fitness may be observed for the adjectives:

All Usual New Incorrect Experimental	results of	(the) present theoretical detailed future	investigation show

Should some of these combinations be forbidden? If so, on what basis? What is the effect of expanding the grammar so that conditional clauses ("if . . ., then . . .") can be generated? One possibility is that adjectives can be semantically classified, so that logically incompatible combinations can be avoided. It is difficult to estimate whether or not such a path of investigation is fruitful.

At the present stage of research, problems of semantic "interference" have occurred infrequently. The generation of certain problem constructions is strongly indicated. For example, the program can be modified so as to produce, at a given point, one hundred sentences in which the construction, "noun of noun of noun," appears; likewise, it can be required that the second of the three nouns be a specific word (e.g., "effect"). From an examination of the output, we hope to gain some insight into the question of semantic compatibility.

REFERENCES

1. Hays, D. G., Basic Principles and Technical Variations in Sentence-Structure Determination, The RAND Corporation, P-1984, May 1960.

2. Edmundson, H. P., K. E. Harper, D. G. Hays and B. J. Scott, "Manual for Postediting Russian Text," Mechanical Translation, Vol. 6, 1961.

3. Kozak, A. S. and C. H. Smith, Studies in Machine Translation--12: A Glossary of Russian Physics, The RAND Corporation, RM-2655, October 1960.

4. Harper, K. E., Measurement of Similarity Between Nouns, The RAND Corporation, RM-4532-PR, May 1965.

5. Sparck-Jones, K., "A Small Semantic Classification Experiment Using Cooccurrence Data," M. L. 196, Cambridge Language Research Unit, January 1967.

6. Harper, K.E., Some Combinatorial Properties of Russian Nouns, The RAND Corporation, RM-5077-PR, September 1966.

SUMMARY - Syntactic and Semantic Problems in Automatic Sentence Generation

A computer program for the automatic generation of sentences was written, based on a simplified dependency grammar for Russian and a vocabulary of 550 words. Each word is provided a list of allowable constituents; the structure of a sentence is conditioned entirely by the lexical items chosen at each node in the dependency tree. Some of the generated sentences exhibit certain syntactic deficiencies (nouns are inadequately complemented or are given the wrong number); other problems result from the "interference" of semantic fields beyond the context of the word-pair.